## Mục lục

# Xấp xỉ một số phân phối xác suất bằng phân phối chuẩn: Ứng dụng và hạn chế

## 1. Đặt vấn đề

Trong lĩnh vực xác suất và thống kê, việc hiểu rõ và mô tả chính xác các phân phối xác suất đóng vai trò quan trọng trong phân tích dữ liệu, ra quyết định và phát triển các mô hình dự đoán. Tuy nhiên, khi đối mặt với các phân phối phức tạp như phân phối nhị thức, phân phối Poisson hay phân phối Chi-bình phương, việc tính toán các xác suất cụ thể hoặc ước lượng các tham số phân phối có thể trở nên rất phức tạp.

Phân phối chuẩn, với tính chất đơn giản và các tính chất toán học dễ áp dụng, đã trở thành một công cụ mạnh mẽ để xấp xỉ các phân phối khác. Điều này không chỉ giúp đơn giản hóa các tính toán mà còn mở ra nhiều ứng dụng thực tiễn trong các lĩnh vực như học máy, kinh tế lượng và phân tích rủi ro.

Mặc dù các định lý kinh điển như Định lý De Moivre-Laplace đã chứng minh sự hiệu quả của xấp xỉ phân phối chuẩn trong nhiều trường hợp [1], câu hỏi đặt ra là khi nào và với điều kiện nào thì xấp xỉ này là hợp lý? Hơn nữa, việc hiểu rõ hơn về những hạn chế và tiềm năng cải tiến của phương pháp này có thể dẫn đến những bước tiến mới trong nghiên cứu và ứng dụng.

Trong bối cảnh đó, bài báo này sẽ trình bày các cơ sở lý thuyết của việc xấp xỉ các phân phối nhị thức, Poisson và Chi-bình phương bằng phân phối chuẩn. Đồng thời, bài viết sẽ đưa ra các số liệu minh họa và đề xuất hướng nghiên cứu tiếp theo để mở rộng và nâng cao tính ứng dụng của xấp xỉ này trong thực tiễn.

## 2. Cơ sở lý thuyết và phương pháp nghiên cứu

### 2.1. Cơ sở lí thuyết

Phân phối chuẩn (Normal Distribution), thường được ký hiệu là $N(\mu,\sigma^2)$ đóng vai trò trung tâm trong lý thuyết xác suất và thống kê. Phân phối này mô tả một tập hợp dữ liệu với hình chuông đặc trưng, có kì vọng bằng $\mu$ và phương sai bằng $\sigma^2$ . Một trong những tính chất quan trọng nhất của phân phối chuẩn là Định lý Giới hạn trung tâm (Central Limit Theorem ), định lý này chỉ ra rằng tổng của một số lượng lớn các biến ngẫu nhiên độc lập, với phân phối giống nhau và có phương sai hữu hạn, sẽ tiệm cận phân phối chuẩn khi kích thước mẫu tăng.

Định lý De Moivre-Laplace là một trường hợp đặc biệt của Định lí giới hạn trung tâm, cho phép xấp xỉ phân phối nhị thức bằng phân phối chuẩn khi số lần thử n đủ lớn. Cụ thể, nếu $X \sim B(n,p)$ (tức là X tuân theo quy luật phân phối nhị thức, với n là số lần thử và p là xác suất xảy ra A trong một lần thử), với kì vọng là $np$ và phương sai là $np(1-p)$, thì khi n đủ lớn, phân phối của X có thể được xấp xỉ bởi phân phối chuẩn: $X \approx N(np; \sqrt{np(1-p)})$.

Tương tự, các phân phối Poisson và Chi-Bình phương cũng có thể được xấp xỉ bằng phân phối chuẩn khi các tham số của chúng đạt đến các giá trị lớn. Với phân phối Poisson $X \sim P(\lambda)$ thì khi $\lambda$ lớn, X có thể được xấp xỉ bởi phân phối chuẩn $X \approx N(\lambda; \sqrt{\lambda})$. Phân phối Chi-Bình phương với k bậc tự do $X \sim \chi^2(k)$ có thể được xấp xỉ bằng $X \approx N(k, \sqrt{2k})$ khi k lớn.

## 2.2. Phương pháp nghiên cứu

a. Chứng minh toán học

Bài báo sẽ xây dựng các chứng minh toán học chi tiết cho việc xấp xỉ phân phối nhị thức, Poisson và Chi-Bình phương bằng phân phối chuẩn [1]. Cụ thể là:

- Chứng minh định lý De Moivre-Laplace: Sử dụng xấp xỉ Stirling và lý thuyết giới hạn để chứng minh phân phối nhị thức có thể được xấp xỉ bằng phân phối chuẩn khi số lần thử lớn.
- Sử dụng Định lý Giới hạn trung tâm để chỉ ra rằng các phân phối Poisson và Chi-Bình phương cũng có thể được xấp xỉ bởi phân phối chuẩn trong những điều kiện nhất định.

b. Mô phỏng và phân tích dữ liệu

Sử dụng các công cụ phần mềm như Python, tiến hành mô phỏng các phân phối nhị thức, Poisson, và Chi-Bình phương với các giá trị tham số khác nhau. Các bước thực hiện bao gồm:

- Tạo mẫu từ các phân phối nhị thức, Poisson, và Chi-bình phương: Chọn các tham số cụ thể và tạo mẫu ngẫu nhiên từ các phân phối này.
- Xấp xỉ bằng phân phối chuẩn: Chuyển đổi các mẫu này thành các biến ngẫu nhiên chuẩn hóa, sau đó so sánh phân phối thực tế của mẫu với phân phối chuẩn tương ứng.
- Phân tích sự khác biệt: Sử dụng các chỉ số như khoảng tin cậy và kiểm định Kolmogorov-Smirnov để đánh giá mức độ tương đồng giữa phân phối thực tế và phân phối chuẩn xấp xỉ.

- Biểu thị kết quả: Kết quả của các phân tích trên sẽ được trình bày dưới dạng bảng số liệu và đồ thị minh họa. Các bảng số liệu sẽ thể hiện các tham số phân phối, giá trị kì vọng và phương sai, cùng với độ lệch chuẩn của các phân phối xấp xỉ. Đồ thị minh họa sẽ trực quan hóa sự so sánh giữa phân phối thực tế và phân phối chuẩn, giúp làm rõ mức độ chính xác của các xấp xỉ trong từng trường hợp cụ thể.

## 3. Kết quả và thảo luận

**Định lý 1.** Cho một biến ngẫu nhiên X có phân phối nhị thức với tham số n và p, tức là $X \sim B(n,p)$. Khi n lớn, phân phối nhị thức có thể được xấp xỉ bởi phân phối chuẩn với kì vọng np và phương sai np(1−p). Cụ thể: với a và b là hai số thực xác định thì $\lim_{n\to\infty} P\left(a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right) = \int_a^b \Phi(x)dx$, trong đó $\Phi(x)$ là hàm phân phối tích lũy của hàm phân phối chuẩn tắc, tức là $\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Định lý này được chứng minh bằng cách sử dụng xấp xỉ Stirling [1], [2].

**Thảo luận**

1. **Phân phối nhị thức**:

- Xấp xỉ phân phối nhị thức bằng phân phối chuẩn cho kết quả khá chính xác. Biểu đồ của phân phối nhị thức và đường cong phân phối chuẩn trùng khớp khá tốt.
- Giá trị kiểm định KS cho thấy sự khác biệt không đáng kể giữa phân phối nhị thức và phân phối chuẩn xấp xỉ.

2. **Phân phối Poisson**:

- Xấp xỉ phân phối Poisson bằng phân phối chuẩn cho thấy một số khác biệt. Biểu đồ của phân phối Poisson và đường cong phân phối chuẩn không trùng khớp hoàn toàn.
- Giá trị kiểm định KS cao cho thấy sự khác biệt đáng kể giữa phân phối Poisson và phân phối chuẩn xấp xỉ. Điều này có thể do $\lambda$ không đủ lớn để xấp xỉ tốt bằng phân phối chuẩn.

3. **Phân phối Chi-Bình hương**:

- Xấp xỉ phân phối Chi-Bình phương bằng phân phối chuẩn có kết quả tốt hơn so với Poisson nhưng vẫn có một số khác biệt.
- Giá trị kiểm định KS cho thấy sự khác biệt giữa phân phối Chi-Bình phương và phân phối chuẩn xấp xỉ là có nhưng không quá lớn.

**Điều kiện áp dụng các xấp xỉ trên:**

- Xấp xỉ phân phối nhị thức bằng phân phối: thường áp dụng khi $np \geq 5; n(1-p) \geq 5$ . Đây là điều kiện để đảm bảo rằng phân phối nhị thức đủ "cân đối" để có thể xấp xỉ bằng phân phối chuẩn.
- Xấp xỉ phân phối Poisson bằng phân phối chuẩn: thường áp dụng khi $\lambda \geq 10$
- Xấp xỉ phân phối Chi-bình phương bằng phân phối chuẩn: thường áp dụng khi số bậc tự do $df > 20$.

## 4. Một số bài toán thực tế:

### 4. 1. Lĩnh vực Mỏ - Địa chất:

*Bài tập 1: Phân tích kích thước hạt*

Một nhà địa chất thu thập kích thước của 50 hạt đá sau khi nổ mìn, các kích thước (mm) được ghi lại như sau: [10, 12, 11, 15, 14, 13, 10, 12, 11, 14, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 10, 12, 11, 14, 16, 15, 13, 14, 12, 11, 15, 17, 19, 18, 16, 14, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33].

1. Tính giá trị trung bình và độ lệch chuẩn của kích thước hạt.

2. Sử dụng phân phối chuẩn để xấp xỉ phân phối kích thước hạt.

3. Tính xác suất để chọn ngẫu nhiên một hạt có kích thước lớn hơn 20 mm.

*Bài tập 2: Đánh giá sai số đo đạc độ sâu*

Độ sâu đo đạc của 30 mỏ khác nhau được ghi lại (m): [100, 105, 102, 98, 110, 107, 101, 104, 106, 108, 95, 97, 99, 103, 111, 109, 98, 105, 102, 100, 101, 106, 104, 107, 99, 103, 111, 108, 98, 97].

1. Tính độ lệch chuẩn của dữ liệu độ sâu.

2. Giả sử dữ liệu tuân theo phân phối chuẩn, hãy tính xác suất để độ sâu của một mỏ nằm trong khoảng từ 100 đến 110 m.

*Bài tập 3: Phân tích áp lực trong tầng chứa dầu*

Áp lực trong tầng chứa dầu được đo tại 40 vị trí khác nhau (psi): [2000, 2020, 2010, 1990, 2050, 2030, 2015, 2025, 2022, 1980, 1995, 2005, 2040, 2025, 2015, 1998, 2012, 2019, 2003, 1985, 2021, 2035, 2009, 2016, 2045, 2001, 2018, 1997, 2038, 2028, 2004, 2011, 2013, 2006, 2055, 2042, 1993, 2026, 2014, 2008, 1994, 2017].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử áp lực tuân theo phân phối chuẩn, hãy tìm xác suất áp lực lớn hơn 2030 psi.

*Bài tập 4: Đánh giá độ bền của mẫu đá*

Kết quả kiểm tra độ bền của 25 mẫu đá (MPa): [50, 55, 52, 54, 51, 53, 56, 57, 58, 55, 52, 51, 50, 54, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Sử dụng phân phối chuẩn để ước lượng xác suất để một mẫu đá có độ bền lớn hơn 60 MPa.

*Bài tập 5: Xác định tuổi của các lớp địa chất*

Tuổi của các lớp địa chất được đo tại 20 điểm khác nhau (triệu năm): [200, 210, 205, 215, 220, 210, 200, 230, 240, 245, 250, 215, 220, 225, 230, 200, 210, 205, 210, 200].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử tuổi lớp địa chất tuân theo phân phối chuẩn, hãy tính xác suất để tuổi của một lớp lớn hơn 225 triệu năm.

### 4. 2. Lĩnh vực kinh tế:

*Bài tập 1: Phân tích mức tiêu thụ năng lượng*

Mức tiêu thụ năng lượng của 50 thiết bị trong khai thác mỏ (kWh): [250, 260, 255, 270, 280, 265, 275, 260, 257, 258, 265, 267, 270, 272, 274, 276, 258, 260, 265, 275, 280, 285, 290, 300, 310, 320, 305, 300, 295, 270, 280, 290, 300, 310, 280, 260, 250, 240, 255, 265, 275, 285, 295, 305, 310, 320, 330, 340, 350, 360].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Sử dụng phân phối chuẩn để xấp xỉ phân phối tiêu thụ năng lượng.

3. Tính xác suất để một thiết bị tiêu thụ năng lượng lớn hơn 300 kWh.

*Bài tập 2: Xác suất vỡ nợ*

Tỷ lệ vỡ nợ của 30 doanh nghiệp khai thác mỏ trong 5 năm qua (%) là: [5, 10, 12, 8, 6, 7, 9, 11, 15, 13, 10, 6, 5, 8, 9, 7, 10, 12, 14, 15, 11, 6, 7, 9, 12, 13, 10, 11, 8, 6].

1. Tính độ lệch chuẩn của tỷ lệ vỡ nợ.

2. Giả sử tỷ lệ vỡ nợ tuân theo phân phối chuẩn, hãy tính xác suất tỷ lệ vỡ nợ lớn hơn 12%.

*Bài tập 3: Phân tích lợi nhuận từ khai thác tài nguyên*

Lợi nhuận từ 40 dự án khai thác tài nguyên (triệu đồng): [10, 12, 11, 15, 14, 13, 10, 12, 11, 14, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 10, 12, 11, 14, 16, 15, 13, 14, 12, 11, 15, 17, 19, 18, 16, 14, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33].

1. Tính giá trị trung bình và độ lệch chuẩn của lợi nhuận.

2. Giả sử lợi nhuận tuân theo phân phối chuẩn, hãy tính xác suất để lợi nhuận từ một dự án lớn hơn 20 triệu đồng.

*Bài tập 4: Phân tích rủi ro trong đầu tư*

Tỷ suất sinh lợi từ 25 khoản đầu tư vào mỏ (%) là: [8, 7, 10, 12, 9, 6, 11, 10, 12, 8, 7, 10, 9, 11, 12, 10, 13, 14, 15, 16, 17, 15, 12, 11, 10].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử tỷ suất sinh lợi tuân theo phân phối chuẩn, hãy tính xác suất để tỷ suất sinh lợi lớn hơn 12%.

*Bài tập 5: Dự đoán lạm phát*

Dự đoán lạm phát trong 10 năm tới (%) là: [2.1, 2.5, 2.3, 2.4, 2.2, 2.6, 2.8, 2.7, 2.9, 3.0].

1. Tính giá trị trung bình và độ lệch chuẩn của dự đoán lạm phát.

2. Giả sử lạm phát tuân theo phân phối chuẩn, hãy tính xác suất lạm phát lớn hơn 2.8%.

### 4.3. Lĩnh vực khoa học công nghệ:

*Bài tập 1: Đánh giá độ tin cậy của thiết bị*

Tuổi thọ của 30 thiết bị khai thác mỏ (năm): [5, 6, 5, 7, 8, 7, 6, 5, 5, 6, 8, 7, 9, 10, 9, 8, 7, 6, 5, 8, 10, 11, 12, 11, 10, 9, 8, 7, 6, 5].

1. Tính giá trị trung bình và độ lệch chuẩn của tuổi thọ thiết bị.

2. Giả sử tuổi thọ tuân theo phân phối chuẩn, hãy tính xác suất để tuổi thọ của một thiết bị lớn hơn 9 năm.

*Bài tập 2: Phân phối thời gian hỏng hóc của cảm biến*

Thời gian hỏng hóc của 25 cảm biến địa chất (tháng): [12, 14, 11, 10, 13, 12, 15, 16, 17, 18, 14, 13, 12, 10, 11, 12, 14, 16, 17, 15, 14, 13, 11, 10, 9].

1. Tính độ lệch chuẩn của thời gian hỏng hóc.

2. Giả sử thời gian hỏng hóc tuân theo phân phối chuẩn, hãy tính xác suất thời gian hỏng hóc lớn hơn 15 tháng.

*Bài tập 3: Dự đoán lưu lượng xử lý*

Lưu lượng xử lý của 20 hệ thống tự động hóa trong khai thác (tấn/giờ): [200, 210, 220, 215, 225, 230, 240, 250, 245, 260, 255, 240, 235, 230, 225, 215, 210, 205, 220, 215].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử lưu lượng xử lý tuân theo phân phối chuẩn, hãy tính xác suất lưu lượng lớn hơn 240 tấn/giờ.

*Bài tập 4: Thời gian đáp ứng của hệ thống quản lý dữ liệu*

Thời gian đáp ứng của 15 hệ thống quản lý dữ liệu (giây): [2.5, 3.0, 2.8, 2.7, 3.2, 3.0, 2.9, 2.6, 2.5, 2.7, 3.0, 3.5, 3.1, 3.2, 2.8].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử thời gian đáp ứng tuân theo phân phối chuẩn, hãy tính xác suất thời gian đáp ứng lớn hơn 3 giây.

*Bài tập 5: Hiệu suất của các thuật toán tối ưu hóa*

Thời gian thực hiện của 30 thuật toán tối ưu hóa (giây): [0.5, 0.6, 0.7, 0.6, 0.5, 0.4, 0.8, 0.7, 0.6, 0.5, 0.9, 0.8, 0.6, 0.7, 0.6, 0.5, 0.4, 0.6, 0.8, 0.9, 1.0, 0.8, 0.6, 0.7, 0.5, 0.4, 0.3, 0.2, 0.6, 0.7].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử thời gian thực hiện tuân theo phân phối chuẩn, hãy tính xác suất thời gian lớn hơn 0.8 giây.

### 4.4. Lĩnh vực môi trường:

*Bài tập 1: Xác suất sự cố tràn dầu*

Tần suất sự cố tràn dầu tại một mỏ dầu trong 5 năm qua (sự cố/năm): [1, 2, 1, 3, 2, 1, 0, 2, 1, 3, 2, 1, 1, 0, 3].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử tần suất tuân theo phân phối chuẩn, hãy tính xác suất tần suất lớn hơn 2 sự cố/năm.

*Bài tập 2: Mức độ ô nhiễm không khí*

Mức độ ô nhiễm không khí (µg/m³) tại 30 điểm quan trắc là: [150, 160, 155, 150, 145, 165, 170, 175, 160, 158, 150, 152, 155, 155, 160, 165, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử mức độ ô nhiễm tuân theo phân phối chuẩn, hãy tính xác suất mức độ ô nhiễm lớn hơn 200 µg/m³.

*Bài tập 3: Xấp xỉ biến động nhiệt độ*

Biến động nhiệt độ trung bình hàng năm tại một khu vực khai thác địa nhiệt (°C) là: [15.0, 15.2, 15.1, 15.3, 15.4, 15.0, 15.2, 15.3, 15.1, 15.4, 15.5, 15.6, 15.3, 15.1, 15.2].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử nhiệt độ tuân theo phân phối chuẩn, hãy tính xác suất nhiệt độ lớn hơn 15.4°C.

*Bài tập 4: Mức độ ô nhiễm nước ngầm*

Mức độ ô nhiễm nước ngầm (mg/L) tại 20 vị trí khác nhau là: [0.1, 0.2, 0.15, 0.3, 0.25, 0.2, 0.15, 0.1, 0.3, 0.35, 0.2, 0.15, 0.25, 0.1, 0.05, 0.2, 0.25, 0.3, 0.15, 0.2].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử mức độ ô nhiễm tuân theo phân phối chuẩn, hãy tính xác suất mức độ ô nhiễm lớn hơn 0.25 mg/L.

*Bài tập 5: Phân tích sự phân bố chất thải*

Khối lượng chất thải (kg) từ 25 khu vực khai thác khoáng sản là: [100, 150, 120, 130, 140, 160, 170, 110, 180, 150, 140, 130, 120, 110, 100, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240].

1. Tính giá trị trung bình và độ lệch chuẩn.

2. Giả sử khối lượng chất thải tuân theo phân phối chuẩn, hãy tính xác suất khối lượng lớn hơn 200 kg.

Những bài tập này không chỉ giúp sinh viên trường Đại học Mỏ - Địa chất hiểu rõ hơn về xấp xỉ phân phối chuẩn mà còn giúp sinh viên áp dụng kiến thức vào thực tiễn trong các lĩnh vực liên quan đến mỏ - địa chất, kinh tế, công nghệ và môi trường.

**Hướng nghiên cứu tiếp theo:**

1. **Thử nghiệm với các tham số khác nhau**: Tiến hành thử nghiệm với các giá trị tham số khác nhau để xem xét mức độ ảnh hưởng của chúng đến sự xấp xỉ phân phối chuẩn.
2. **Sử dụng các phương pháp xấp xỉ khác**: Áp dụng các phương pháp xấp xỉ khác như xấp xỉ Gamma hoặc xấp xỉ Poisson bằng phân phối siêu bội để so sánh hiệu quả.
3. **Phân tích sự ảnh hưởng của kích thước mẫu**: Nghiên cứu sự ảnh hưởng của kích thước mẫu đến kết quả xấp xỉ và độ chính xác của các xấp xỉ này.
4. **Mô phỏng thêm các phân phối khác**: Tiến hành mô phỏng và xấp xỉ các phân phối khác như phân phối Beta, phân phối Gama, và phân phối Student’s t để mở rộng phạm vi nghiên cứu.

## Kết luận

Bài báo đề cập đến khả năng xấp xỉ phân phối chuẩn đối với các phân phối khác như phân phối nhị thức, phân phối Poisson và phân phối Chi-bình phương, từ nền tảng lý thuyết đến ứng dụng thực tiễn. Các kết quả phân tích cho thấy phân phối chuẩn, với những đặc tính toán học rõ ràng và dễ áp dụng, đóng vai trò quan trọng trong việc đơn giản hóa các tính toán phức tạp liên quan đến các phân phối khác. Bài viết cũng đã cung cấp những ví dụ minh họa và biểu đồ trực quan, chứng minh hiệu quả của phương pháp xấp xỉ này trong thực tiễn. Tuy nhiên, bài báo cũng nhận diện những hạn chế của phương pháp xấp xỉ phân phối chuẩn, đặc biệt là khi các điều kiện cần thiết không được đảm bảo. Việc hiểu rõ hơn về những hạn chế này sẽ là nền tảng cho các nghiên cứu tiếp theo nhằm cải thiện độ chính xác và mở rộng ứng dụng hơn.

## Tài liệu tham khảo


1. Grinstead and Snell's (2006). *Introduction to Probability,* New York: Wiley.
2. Johnson, N. L., & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*. New York: Wiley.
3. Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics and Probability for Engineers*. Hoboken, NJ: Wiley.